What is multivariate data set?

- A data set in which several variables are measured on each sampled unit.
- We have n units (individuals) and p variables in a multivariate data set where $p > 1$.
- Organization of data: x_{ij} : measurement of the $j-$th variable on the $i-$th unit.
- Observations can be displayed in form (collection of numbers with n rows and p columns):

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

or use $\{x_{ij}\}$.

- The rows $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$ denote the $i-$th observation of p-dimensional space.

Descriptive Statistics

- Mean vector (center of the n observations in $\mathbb{R}^p$):

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

- Sample covariance matrix (variation of n observations in $\mathbb{R}^p$):

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

  - Diagonal entries are the sample variances of the p variables
  - Off-diagonal entries are the sample covariances between two of the variables

- Sample correlation matrix: $\mathbf{R} = \{R_{ij}\}$, where $R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$
  - Diagonal entries are 1's
  - Off-diagonal entries are the respective sample correlation values.

Visualization

- One-variable at a time: using histograms, box plots, etc.
- Graphics that allow us at look at several variables at once are more useful
  - 2-d Scatter plot
  - $p > 2$, Scatter plot matrix $p \times p$ array of 2-d scatter plots: useful to study pairwise relationship
  - conditioning (scatter) plots: scatter plots of two variables conditioning on the third variable. useful when the third variable is grouping variable.

Visualization

- Chernoff-Flury Faces
  - Represent p variables using p specific characteristics of the face
  - eye size, eye brow, hair line, etc...
- Andrew's Curve: multivariate observation $X = (X_1, \ldots, X_p)$ is transformed into a curve.

$$f(t) = \left\{ \begin{array}{ll} \frac{X_1}{\sqrt{2}} + X_2 \sin(t) + X_3 \cos(t) + \cdots + X_{p-1} \sin(\frac{p-1}{2}t) + X_p \cos(\frac{p-1}{2}t) & \text{for odd } p \\ \frac{X_1}{\sqrt{2}} + X_2 \sin(t) + X_3 \cos(t) + \cdots + X_{p-1} \sin(\frac{p}{2}t) & \text{for even } p \end{array} \right. .$$

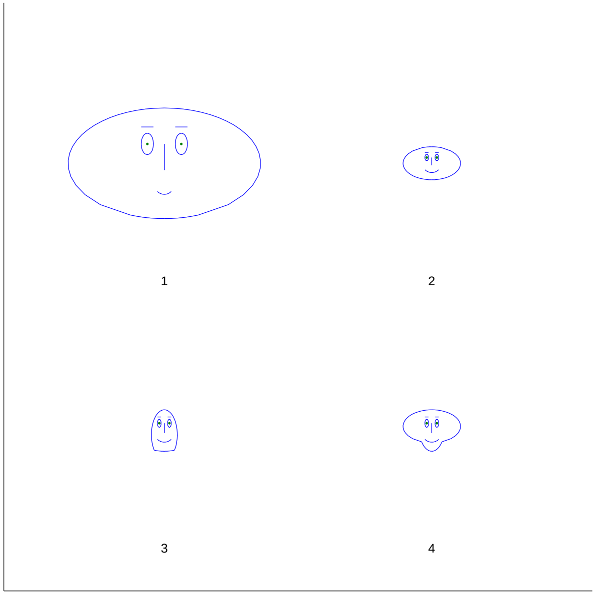
  - Variable reordering?
- Parallel Coordinate Plots:
  - Plot observations not in an orthogonal coordinate system, but in a parallel axes coordinate system.
  - Works best when n is small to moderate.
  - Variable reordering?

Toy Example

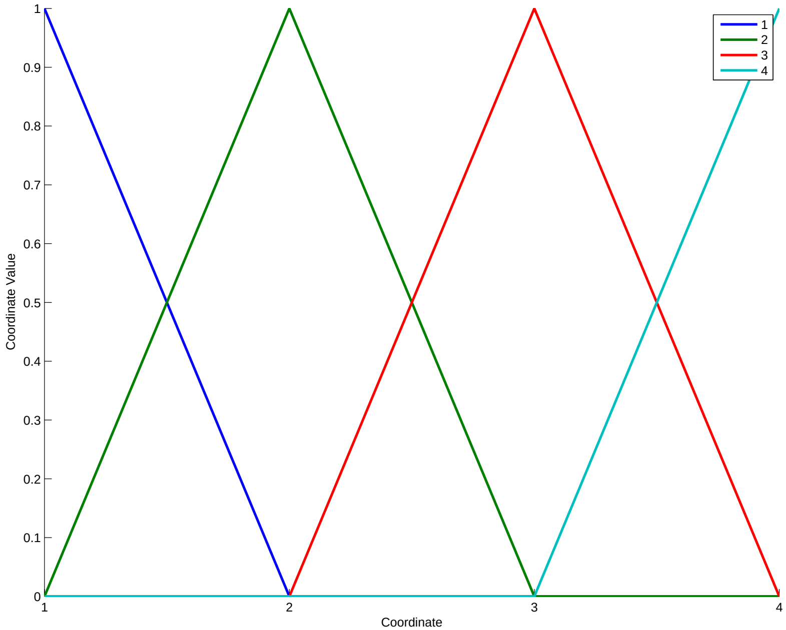
- $\mathbf{x}_1 = (1, 0, 0, 0)$, $\mathbf{x}_2 = (0, 1, 0, 0)$, $\mathbf{x}_3 = (0, 0, 1, 0)$, $\mathbf{x}_4 = (0, 0, 0, 1)$

Andrew's Curve

Face plot



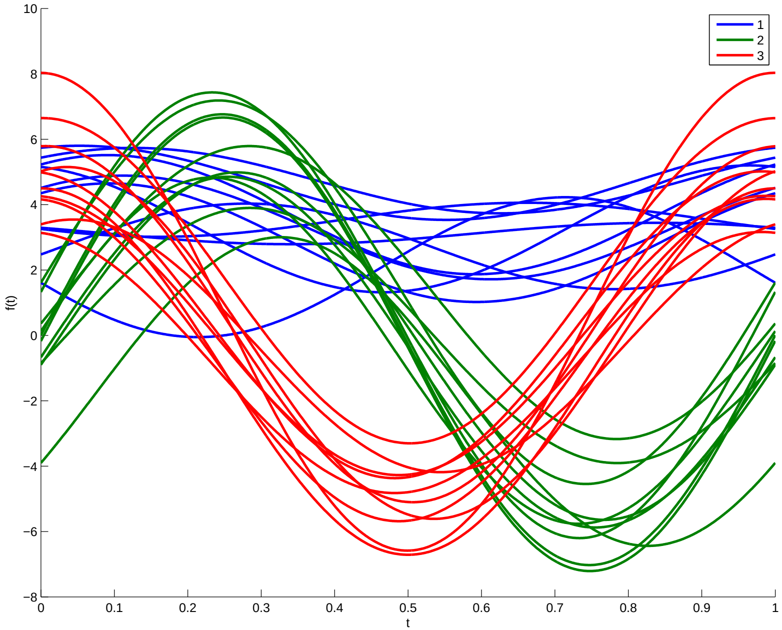
Parallel coordinate plot



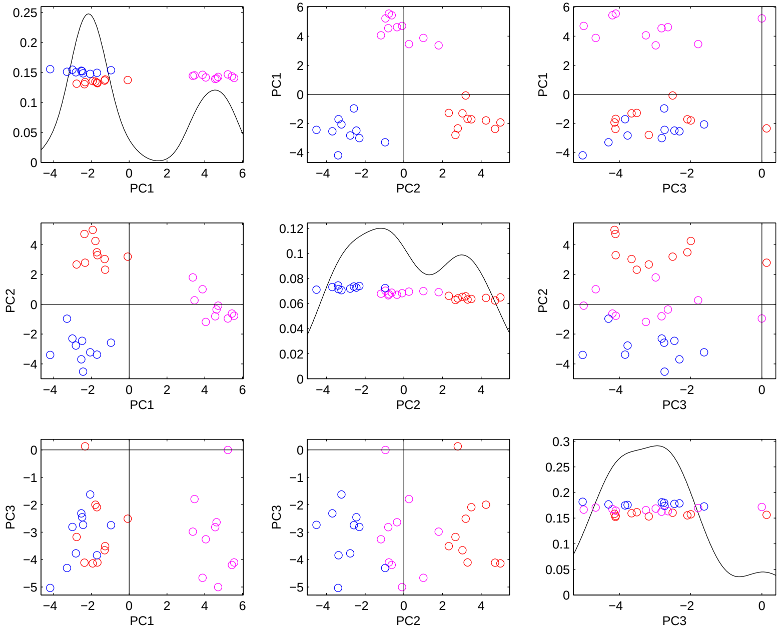
Toy Example

- Group 1 $X \sim N(\mu, \mathbf{I}), \mu = (5, 0, 0)$
- Group 2 $X \sim N(\mu, \mathbf{I}), \mu = (0, 5, 0)$
- Group 3 $X \sim N(\mu, \mathbf{I}), \mu = (0, 0, 5)$

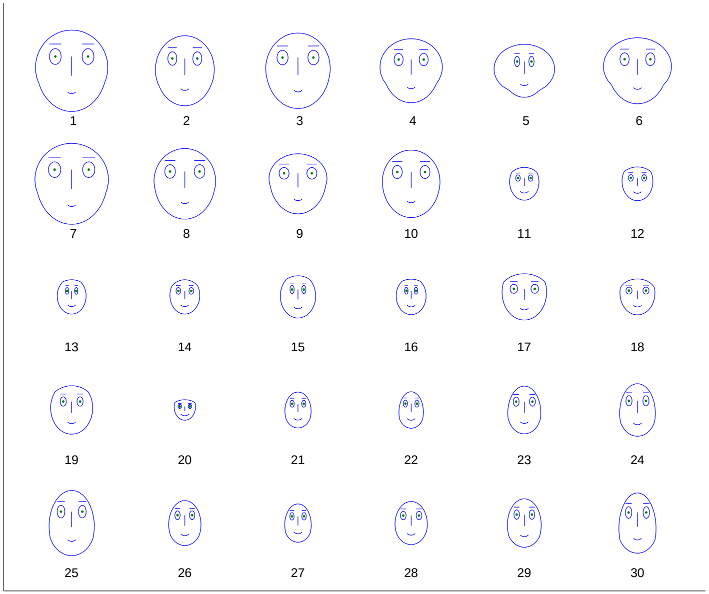
Andrew's Curve



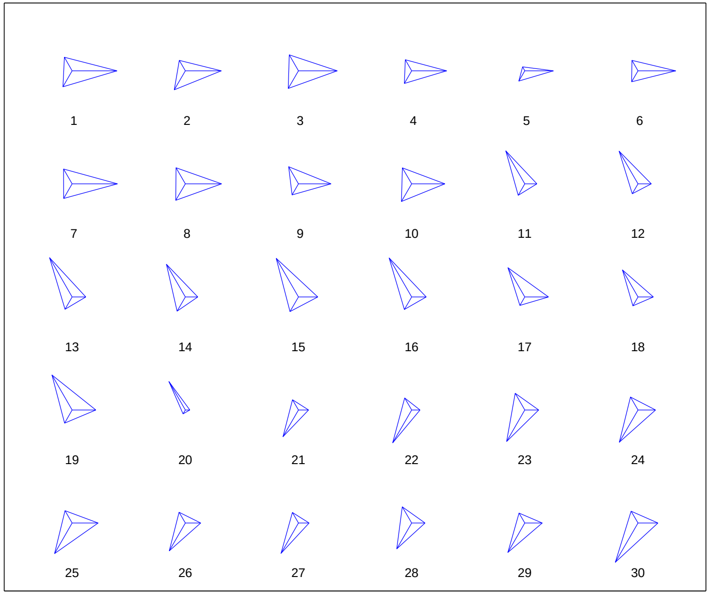
PCA scatter plot (Conditioning plot)



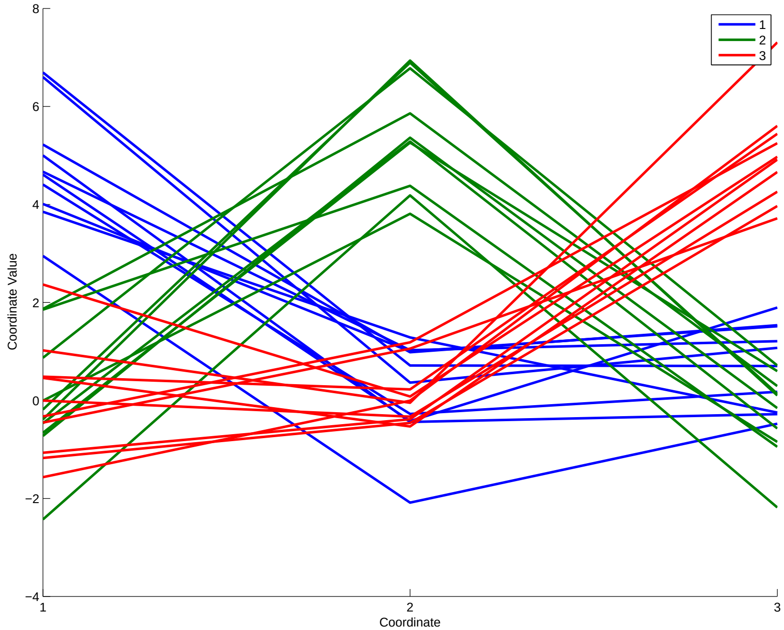
Face plot


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Star plot


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

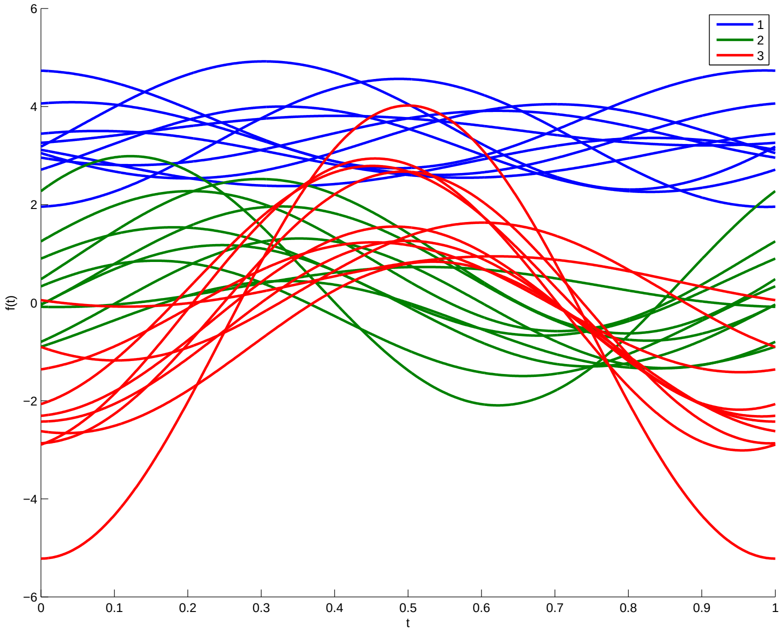
Parallel coordinate plot



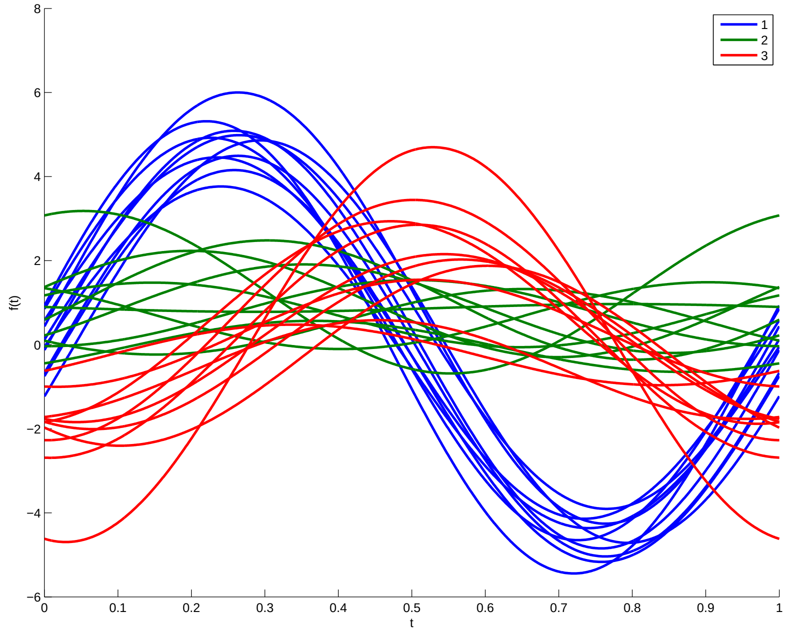
Reordering of variables?

- Case 1
  - Group 1 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (5, 0, 0)$
  - Group 2 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, 1, 0)$
  - Group 3 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, 0, -2)$
- Case 2
  - Group 1 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, 5, 0$
  - Group 2 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (1, 0, 0)$
  - Group 3 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, 0, -2)$
- Case 3
  - Group 1 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, 0, 5)$
  - Group 2 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (1, 0, 0)$
  - Group 3 $X \sim N(\mu, .7^2\mathbf{I}), \mu = (0, -2, 0)$

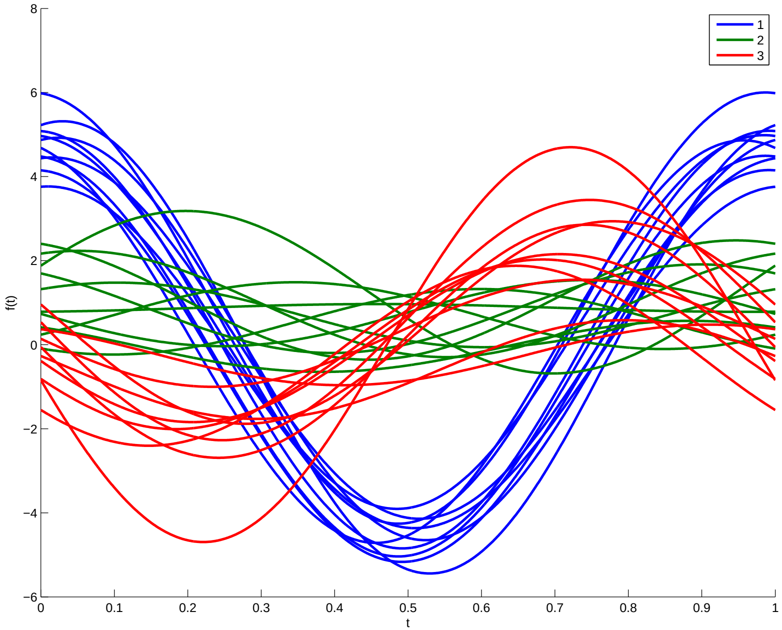
Andrew's Curve: Case 1



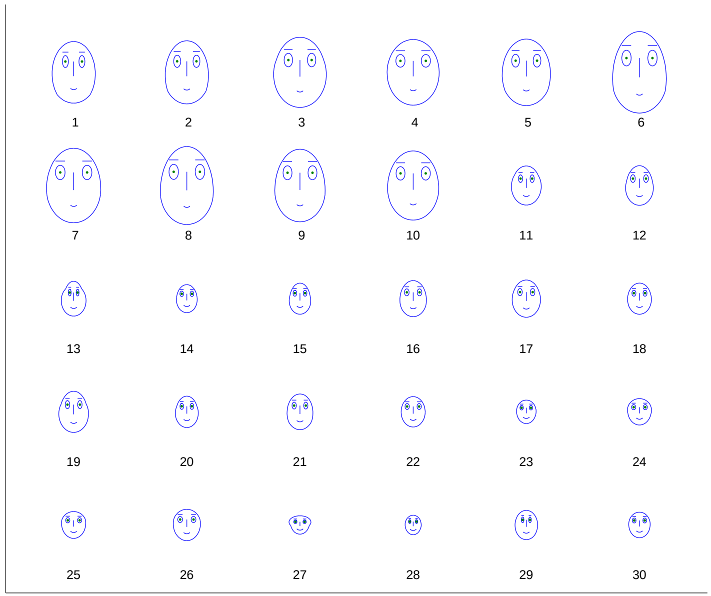
Andrew's Curve: Case 2



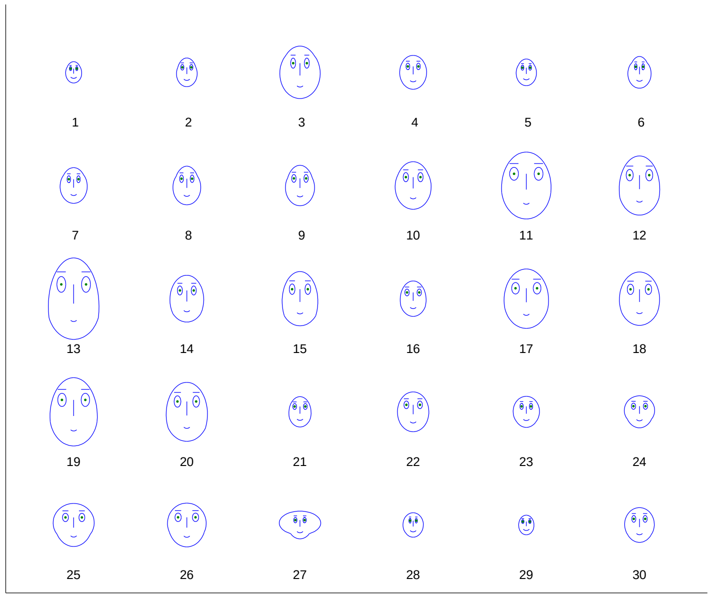
Andrew's Curve: Case 3



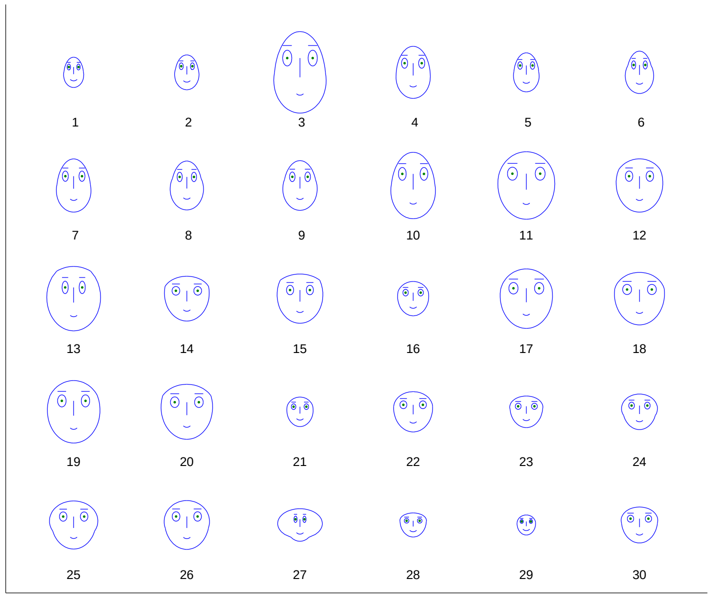
Face plot: Case 1



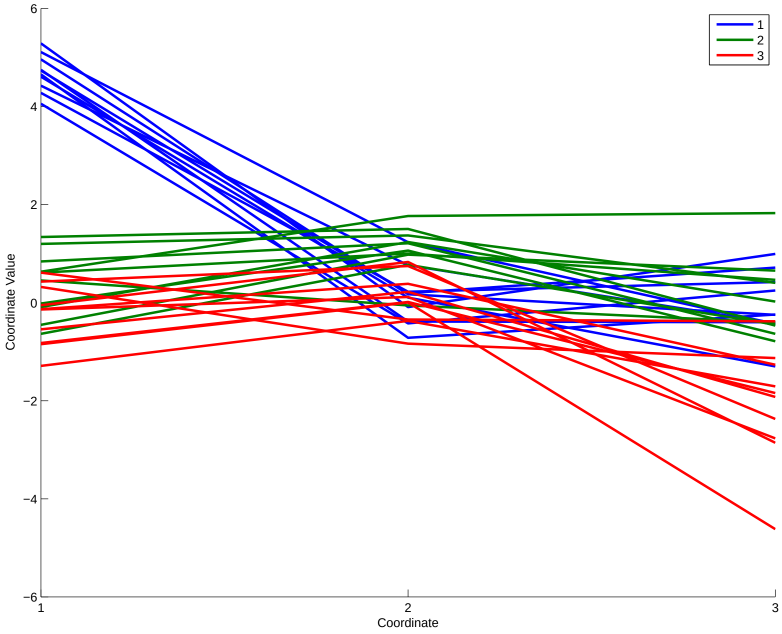
Face plot: Case 2



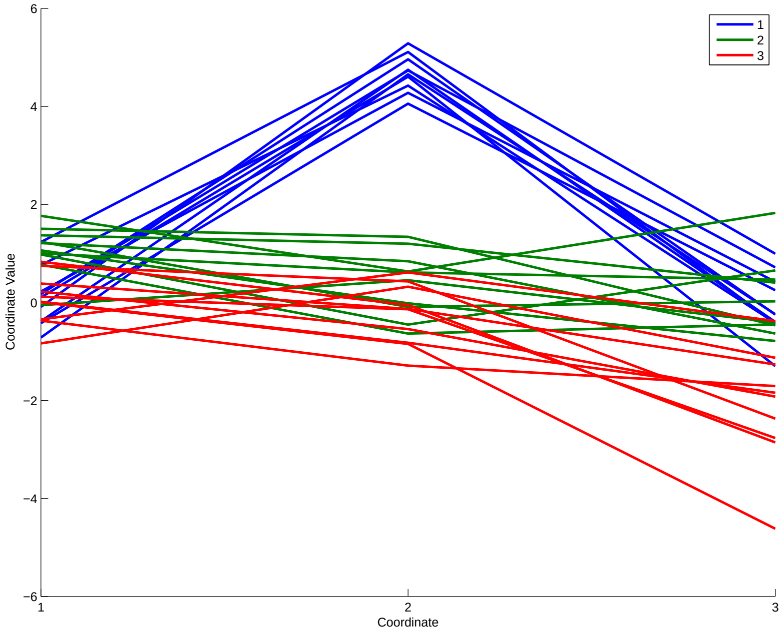
Face plot: Case 3



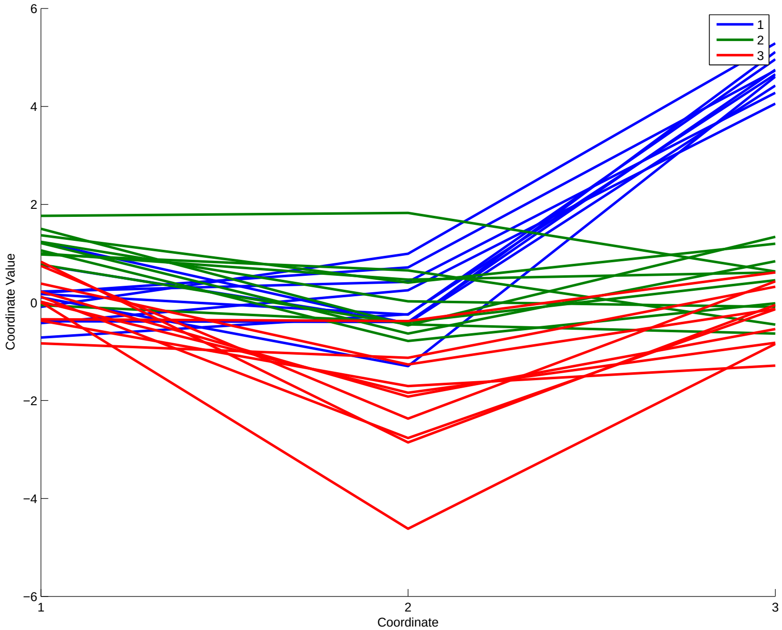
Parallel coordinate plot: Case 1



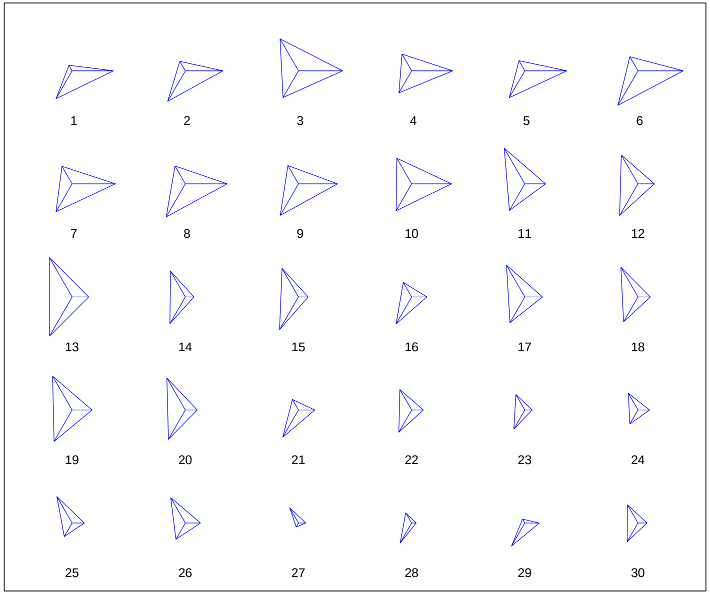
Parallel coordinate plot: Case 2



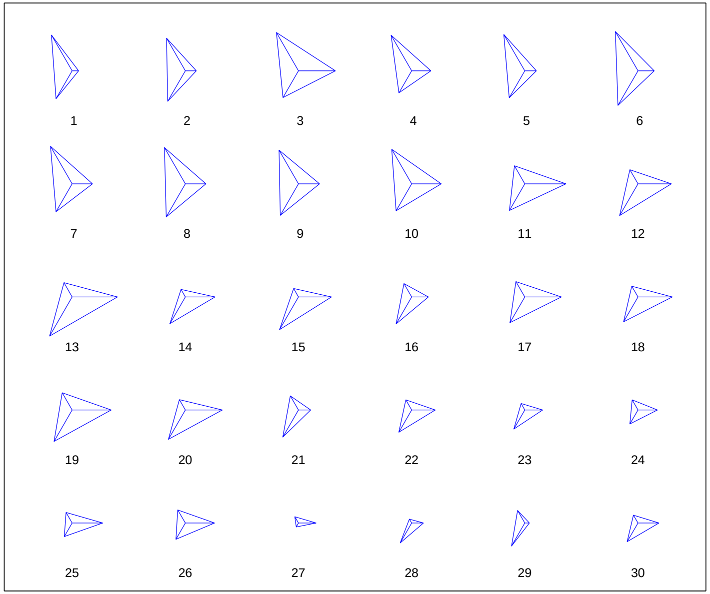
Parallel coordinate plot: Case 3



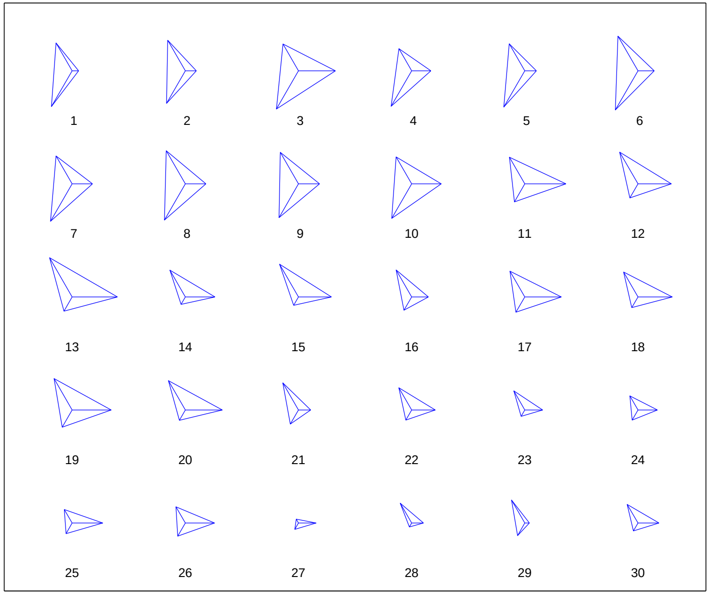
Star plot: Case 1



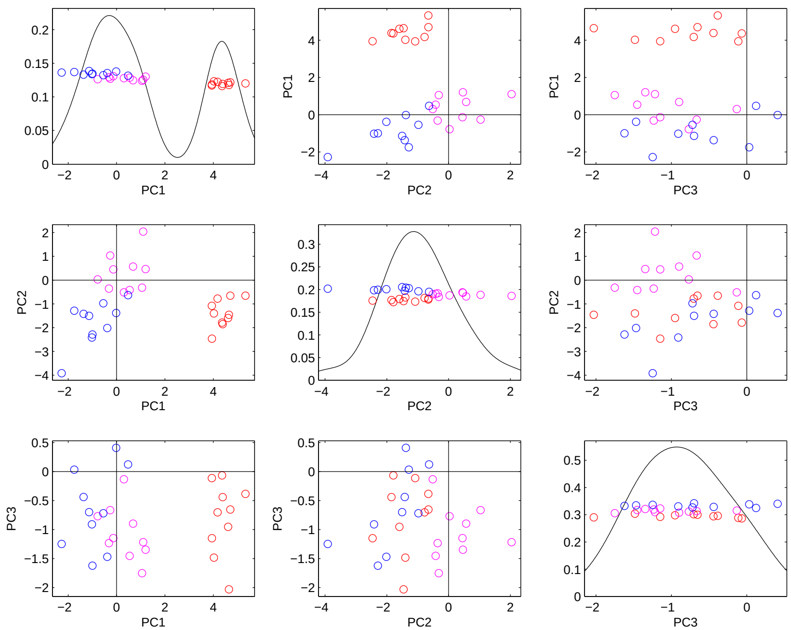
Star plot: Case 2



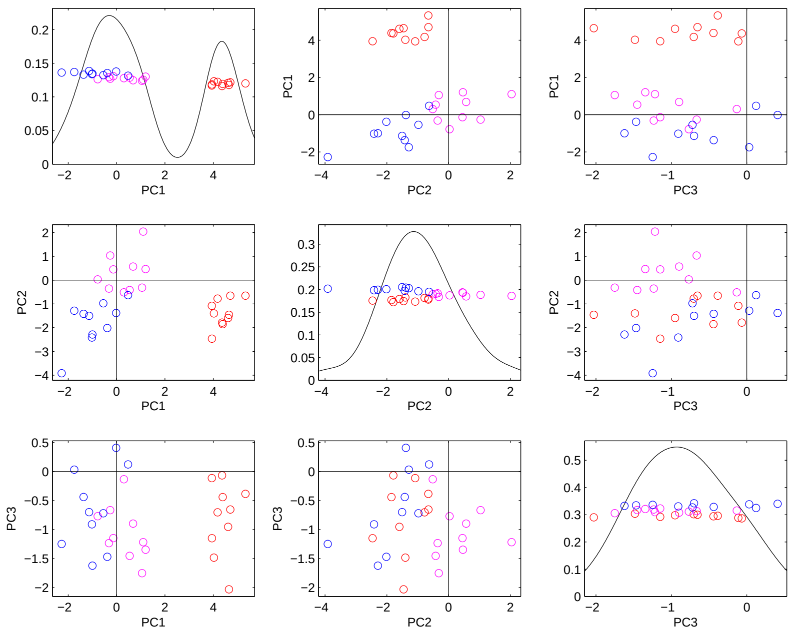
Star plot: Case 3



PCA scatter plot: Case 1



PCA scatter plot: Case 2



Visualization Summary

- Scatter plot matrix: useful to study pairwise relationship. different orthogonal axes coordinate system can be used.
- Andrew's Curve: useful to detect subgroups. Order of the variables plays an important role.
- Face plot: useful to detect subgroups. Order of the variables plays an important role.
- Parallel coordinate plot: useful to detect subgroups